Universidade de Brasília

Modelo de regressão pertencente à família Weibull com fração de cura

Thuany de Aguiar Santos

Brasília
2013

Thuany de Aguiar Santos
Bacharel em Estatística

Modelo de regressão pertencente à família Weibull com fração de cura

Orientadora:
Profa. Dra. **Juliana Betini Fachini**

Brasília
2013

DEDICATÓRIA

À Deus, que é a minha força e esperança em todos os momentos.

Aos meus pais,

Alda Aguiar *e* ***Valdir Santos****, por todo amor e ensinamento.*

AGRADECIMENTOS

A Deus que iluminou meu caminho e guiou meus passos .

A Professora **Juliana Betini Fachini** pela dedicação e comprometimento ao longo desse trabalho, e acima de tudo pela paciência e generosidade.

A todos os professores da UNB que contribuiram para a minha formação.

Aos meus amigos **Marcus Vinicius, Camila Leal e Barbara Lopes** pelas muitas horas de trabalho em conjunto, sugestões, orientações e, principalmente, paciência e motivação.

A minha mãe **Alda** pelo amor e apoio incondicional e pelas orações di´arias.

A minha irmã **Airina** pelo cuidado e companherismo.

Ao **Valdir e Josane** por todo o suporte e apoio.

Aos meus amados tios **Valmir e Estela**, pelo apoio durante a graduação e por me receberem de braços abertos em seu lar.

Um agradecimento especial ao **Gabriel Junqueira** por todo carinho, apoio, cuidado e amor.

Aos colegas de trabalho, em especial **Vanessa, Lais e Lucelia** que nas horas que precisei estiveram presentes.

Aos colegas de faculdade que ao longo da graduação contribuiram de alguma forma para minha formação.

SUMÁRIO

RESUMO

Modelo de regressão pertencente à família Weibull com fração de cura

Neste trabalho são reunidos alguns modelos da família Weibull (Weibull, Weibull Modificada, Weibull exponencializada e Weibull modificada generalizada) e técnicas para estudo de situações em análise de sobrevivência. Foi proposto o modelo de regressão Weibull modificada generalizada com fração de cura, pois, esse modelo considera a presença de indivíduos curados e é generalização dos demais. Os parâmetros dos modelos foram estimados pelo método de máxima verossimilhaça sujeito a restrição nos espaço paramétrico. Um conjunto conjunto de dados médicos foi utilizado para a análise e a aplicação dos modelos propostos.

Palavras-chave: Modelos de regressão; Análise de sobrevivência, Dados censurados; Verossimilhança restrita; Fraçao de cura; Família Weibull

1 INTRODUÇÃO

A análise de sobrevivência é uma área da Estatística que avalia o tempo decorrido até a ocorrência de um evento ou situação de interesse e se caracteriza por utilizar a informação de todos os indivíduos presentes no estudo, inclusive daqueles em que as observações estão incompletas. Nesse caso, os dados são chamados de observações censuradas. Um exemplo desse tipo de observação, acontece em estudos médicos em que o paciente morre por outras causas não relacionadas ao evento de interesse, ou muda de cidade não dando continuidade ao tratamento.

Essa metodologia pode ser utilizada em diversas áreas do conhecimento. Entretanto, é particularmente importante em pesquisas de saúde, para estudos em que geralmente o evento de interesse é a morte de um paciente ou a recidiva de uma doença. Na engenharia ela também é muito utilizada, porém, é conhecida por análise de confiabilidade. Nesse caso geralmente estuda-se o tempo de vida ou a durabilidade de alguns produtos em geral.

Uma característica que pode ocorrer em dados de sobrevivência é a existência de indivíduos não suscetíveis ao evento de interesse, esses indivíduos são chamados curados ou imunes, pois nunca vão experimentar o evento estudado. Quando existe fração de cura o modelo estima a fração de sobrevivência (Berkson e Gage, 1952).

Assim como em outras áreas, na análise de sobrevivência pode-se estudar o comportamento do tempo quando inclue-se outras covariáveis na análise. Pois, a variável tempo pode estar sendo influenciada por outras covariáveis da população estudada. Usa-se então modelos de regressão para englobar as covariáveis no modelo.

Ao se considerar todas as características (tempo de falha, censura, covariáveis e fração de cura), este trabalho tem como objetivo principal propor um modelo de regressão paramétrico com fração de cura em que os tempos assumem distribuições da família Weibull, para um conjunto de dados médicos apresentados por Ibrahim et al (2001).

O software que será utilizado para o desenvolvimento das análises estatísticas da aplicação desse trabalho é o software estatístico R versão 2.13.2.

2 REVISÃO DE LITERATURA

2.1 Notação e conceitos básicos

Para fundamentar o trabalho que será desenvolvido na área de análise de sobrevivência, serão expostos, a seguir, conceitos e notações utilizadas na literatura.

2.1.1 Tempo de falha

Os dados de análise de sobrevivência têm como característica fundamental a presença do tempo de falha, ou seja, o tempo até que o evento de interesse aconteça. Esse tempo é o intervalo entre o tempo inicial do estudo até o tempo de ocorrência da falha e é constituído por três elementos (COLOSIMO e GIOLO, 2006):

- O tempo inicial, tempo que marca o início do estudo;
- A escala de medida (geralmente utiliza-se o tempo real) ;
- O evento de interesse, chamado de falha. Entende-se por falha o acontecimento do evento de interesse definido no estudo.

Em pesquisas clínicas, as falhas geralmente são a morte ou a recidiva de certa doença. Na engenharia, geralmente fazem-se testes com itens mecânicos ou componentes eletrônicos para obter informação de sua durabilidade. Atualmente, tem-se utilizado a análise de sobrevivência também na área financeira, em que o evento de interesse na maioria das vezes é a inadimplência do cliente.

2.1.2 Censura

A análise de sobrevivência se diferencia dos métodos clássicos, pois leva em consideração os dados de todos os indíviduos da análise, inclusive os que não têm a informação do tempo de ocorrência do evento de interesse (esses são chamados de dados censurados).

Os dados censurados podem ser classificados de diferentes formas : dados censurados à direita, à esquerda ou intervalar.

Os tipos de censura à direita são assim denominados porque o tempo de ocorrência do evento de interesse está à direita do tempo registrado. A censura à direita tem três subdivisões:

- Censura do tipo I: O término do estudo acontece em um tempo t_0, estabelecido antes do inicio do estudo, e ao final do estudo algumas observações não falharam. Um exemplo deste tipo de censura são testes de durabilidade de uma lâmpada. No tempo determinado o estudo termina e algumas lâmpadas podem ainda não ter falhado.

- Censura do tipo II: o estudo termina quando um número pré-estabelecido de indivíduos falham. Em estudos biológicos em que geralmente não há tempo nem recursos para esperar que todos os indivíduos falhem, pode-se estabelecer que o fim do estudo será atingido quando uma quantidade pré-determinada de indivíduos falharem.

- Censura do tipo aleatório: quando um indivíduo é retirado no decorrer do estudo sem que a falha tenha ocorrido. Um exemplo clássico desse tipo de censura são estudos clínicos em que alguns indivíduos morrem por causas diferentes do evento de estudo.

O mecanismo de censura à esquerda ocorre quando o evento de interesse já aconteceu antes do tempo observado. Um exemplo desse mecanismo é quando se pesquisa a reincidência de uma doença, e dentre os pacientes analisados existem aqueles em que a doença reincidiu antes do tempo de registro.

O terceiro mecanismo de censura é a censura intervalar, que acontece devido ao tempo de falha exato ser desconhecido sendo, no entanto, o intervalo de tempo em que a falha ocorre conhecido. Por exemplo, em estudos que são realizados com visitas periódicas, não se sabe com exatidão quando o evento de interesse aconteceu, mas apenas o intervalo de tempo em que ocorreu (COLOSIMO e GIOLO, 2006).

2.1.3 Representando o tempo de sobrevivência

O tempo de sobrevivência de um indivíduo é uma variável aleatória não negativa T que essa variável representa o tempo de falha e tem uma distribuição de probabilidade associada a ela. Em termos probabilísticos, suponha que a variável T tem uma função densidade de probabilidade f(t) que satisfaz as seguintes proposições (Magalhães,2011):

1. $f(t) \geq 0, \forall t \in \mathbb{R};$

2. $\int_{\infty}^{-\infty} f(t)dt = 1.$

A função de distribuição acumulada da T é definida por (Magalhães, 2011):

$$P(T \leq t) = F(t) = \int_{-\infty}^{t} f(w) dw \forall t \in \mathbb{R}. \quad (1)$$

Existem duas funções muito usadas para representar o tempo de sobrevivência que estão diretamente relacionadas com a função de probabilidade. São elas: a função de sobrevivência e a função risco. A função de sobrevivência representa a probabilidade de uma observação sobreviver ao tempo t e é definida por (COLOSIMO e GIOLO, 2006):

$$S(t) = P(T \geq t).$$

Essa função pode ser escrita em termos da função de distribuição acumulada, S(t)=1-F(t), em que F(t) representa a probabilidade da observação não sobreviver a um tempo t. A função de sobrevivência tem algumas propriedades básicas, entre elas o fato dessa função no tempo zero ser igual a 1 e quando o tempo (t) tende a infinito, a função de sobrevivência se aproximar de zero. Outra característica fundamental é que essa função é não crescente.

Outra função muito utilizada na análise de sobrevivência é a função risco ou taxa de falha. Essa função representa o limite da probabilidade do indivíduo falhar em um intervalo de tempo $(t, t + \Delta t)$ dado que ele sobreviveu até o tempo t. A função é definida por (COLOSIMO e GIOLO, 2006):

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \triangle t | T \geq t)}{\triangle t}.$$

As funções de sobrevivência e risco são relacionadas pela seguinte fórmula (COLOSIMO e GIOLO, 2006):

$$\lambda(t) = \frac{f(t)}{S(t)}$$

e

$$S(t) = \exp\left[- \int_{0}^{t} \lambda(u) du \right]$$

Portanto, a função risco é a divisão entre a função densidade de probabilidade e a função de sobrevivência.

2.1.4 Estimador de Kaplan Meier

Quando se faz um estudo de análise de sobrevivência, a análise descritiva dos dados é uma etapa fundamental para o desenvolvimento do estudo. Devido à existência da censura nos dados de sobrevivência, as metodologias tradicionais da estatística descritiva não são adequadas. Neste caso, para realizar o estudo descritivo dos dados, usa-se como componente fundamental a função de sobrevivência. A partir desta função, estima-se as demais estatísticas de interesse. Para estimar essa função, existem alguns estimadores não paramétricos bastante conhecidos, entre eles o estimador de Kaplan-Meier. A construção desse estimador envolve uma sequência de passos, levando-se em consideração a informação do passo anterior para a obtenção do passo seguinte. As considerações preliminares que devem ser feitas para a utilização do estimador de Kaplan-Meier são (COLOSIMO e GIOLO, 2006):

- $t_1 < t_2 < ... < t_k$ são os k tempos distintos e ordenados de falha;
- d_j é o número de falhas em t_j, $j = 1, ..., k$ e
- n_j é o numero de indivíduos sob risco (isto é, que não falharam) em t_j.

$$\widehat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right). \tag{2}$$

O estimador de Kaplan-Meier (5) é um estimador não paramétrico de máxima verossimilhança para a função de sobrevivência. Quando existe censura do tipo I e II o estimador se mantém mas a função de sobrevivência estimada pode nunca assumir o valor zero, pois as últimas observações podem ser censuradas. As propriedades básicas desse estimador são (COLOSIMO e GIOLO, 2006):

- não-viciado para amostras grandes;
- fracamente consistente ;
- converge assintoticamente para um processo gaussiano e
- é o estimador não paramétrico de máxima verossimilhança de $S(t)$.

Algumas estimações de quantidades básicas, como estimativas de percentis, tempo mediano e tempo médio, são obtidas a partir da função de sobrevivência estimada pelo método de Kaplan-Meier. A estimação do tempo médio é obtida calculando-se a área sob a curva de Kaplan-Meier estimada. Porém, se o maior tempo observado for uma censura, a função de sobrevivência não assume o valor zero o que inviabiliza a obtenção do tempo médio e sugere-se usar o tempo mediano como medida de posição da distribuição do tempo de vida.

2.1.5 Determinação empírica da forma da função de risco

Depois da análise não paramétrica que tem o intuito de obter informações preliminares dos dados que ajudarão na escolha do modelo, é necessário descobrir qual função distribuição de probabilidade modela melhor esses dados. Uma metodologia que ajuda a nortear a escolha de algumas distribuições que possam se adequar aos dados é a curva do tempo total em teste (curva TTT) (Aarset, 1987). A curva TTT é o gráfico da função G(r/n) versus r/n sendo r=1,...,n e $T_{i:n}$, i=1,...,n as estatísticas de ordem da amostra.

$$G(r/n) = \frac{[\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}]}{\sum_{i=1}^{n} T_{i:n}}.$$

A curva gerada a partir desse gráfico pode ter várias formas e cada uma delas está associada a uma forma diferente da função taxa de falha.

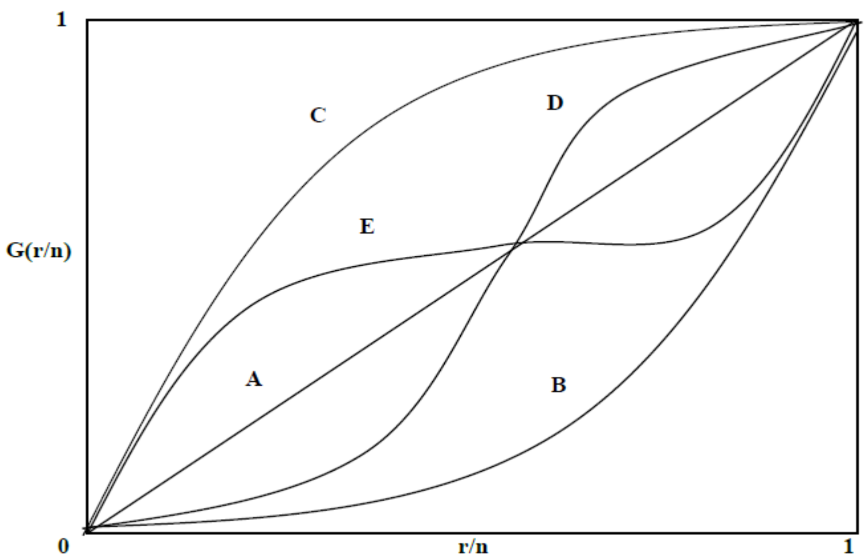

Figura 1 – Gráficos ilustrativos de algumas curvas TTT.

A Figura 1 permite a visualização de algumas possíveis formas da curva TTT. As relações entre essas formas e a forma da função risco são:

- A reta diagonal representada pela curva A está associada a uma função risco constante.
- A curva B que tem a forma convexa é relacionada a função risco descrescente.
- A curva C, côncava, se associa a função taxa de falha crescente.
- Quando a curva é convexa e depois côncava, curva D, a função taxa de falha tem a forma de banheira .
- Se a curva é primeiramente côncava e depois convexa, a função risco é unimodal.

Portanto, através da curva TTT descobre-se a forma da função risco, o que restringe as funções de probabilidade que melhor se adequarão aos dados. Por este motivo, essa metodologia não paramétrica é muito utilizada como análise descritiva dos dados.

2.2 Distribuições de Probabilidade

As distribuições de probabilidade tem como objetivo descrever uma variável aleatória com ela é possível calcular a probabilidade de ocorrência intervalos de valores dessa variável. Existem na literatura várias distribuições de probabilidade que têm particularidades diferentes, o que torna cada distribuição mais adequada a um tipo diferente de variável aleatória.

Na análise de sobrevivência, existe a particularidade da variável resposta ser não negativa, pois o tempo não pode assumir valores negativos. Devido a essa particularidade, as distribuições de probabilidade que conseguem modelar os dados de sobrevivência são distribuições em que a variável aleatória é definida para valores maiores ou iguais a zero.

Com o intuito de modelar os dados desse estudo, serão testadas distribuições da família Weibull para um modelo de regressão e tentar descobrir qual dos modelos melhor incorpora as particularidades dos dados.

2.2.1 Distribuição Weibull

A distribuição Weibull (COLOSIMO e GIOLO, 2006) é uma das funções mais utilizadas na análise de sobrevivência, especialmente para dados na área da saúde e engenharia. A distribuição Exponencial é um caso particular da Weibull. A função densidade de probabilidade e a função de sobrevivência Weibull são descritas como:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^\gamma \right\}, t \geq 0,$$

e

$$S(t) = \exp\left\{ -\left(\frac{t}{\alpha}\right)^\gamma \right\}.$$

A distribuição Weibull tem a função taxa de falha definida por:

$$h(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}, \tag{3}$$

em que $t \geq 0$, $\alpha \geq 0$ parâmetro de escala e $\gamma \geq 0$ parâmetro de forma. Uma das particularidades da distribuição Weibull é que ela tem a distribuição exponencial como caso particular quando $\gamma = 1$ e a função risco constante devido à propriedade de falta de memória da distribuição exponencial. A função taxa de falha definida em (3), assume um comportamento crescente quando $\gamma > 1$ e um comportamento descrescente quando $\gamma < 1$.

2.2.2 Distribuição Weibull modificada

Proposta por Lai, Min XIE e Murthy(2003), a distribuição Weibull modificada (DWM) que modela uma variável aleatória não negativa T tem as seguintes funções de densidade de probabilidade, função acumulada de probabilidade e função de sobrevivência (Carrasco,2007):

$$f(t) = \alpha(\gamma + \lambda t)t^{\gamma-1} \exp(\lambda t) \exp\{-\alpha t^\gamma \exp(\lambda t)\}, (4)$$

$$F(t) = 1 - \exp\{-\alpha t^\gamma \exp(\lambda t)\},$$

$$S(t) = \exp\{-\alpha t^\gamma \exp(\lambda t)\},$$

em que $\alpha > 0$ é o parâmetro de escala, $\gamma \geq 0$ é o parâmetro de forma da distribuição, e o fator λt é o fator acelerador na sobrevida. Portanto, quando o tempo aumenta, λ opera como fator de fragilidade na sobrevida. A forma que a função risco assume depende apenas do fator γ, porque $t^{\gamma-1}$ e os dois parâmetros restantes não tem influência. A função risco associada a essa distribuição é definida por:

$$h(t) = \alpha(\gamma + \lambda t)t^{\gamma-1} \exp(\lambda t). \tag{5}$$

Essa distribuição é de grande utilidade, pois apresenta diversas formas da função taxa de falha. Quando $\gamma \geq 1$, a função assume uma forma crescente, e se $0 < \gamma < 1$, tem forma de banheira (inicialmente decresce e depois cresce).

A função Weibull modificada tem relação com outras distribuições de probabilidade e apresenta algumas distribuições como casos particulares:

1. Quando $\lambda = 0$, a distribuição Weibull modificada se reduz à forma da distribuição Weibull. Portanto, a distribuição Weibull e a distribuição exponencial são casos particulares da distribuição Weibull modificada.

2. Para $\gamma = 0$, a distribuição Weibull modficada assume a forma da distribuição do Valor extremo tipo I: $S(t) = \exp\{-\alpha \exp(\lambda t)\}$.

Isto mostra a flexibilidade da distribuição de probabilidade Weibull modificada.

2.2.3 Distribuição Weibull exponencializada

A distribuição Weibull exponencializada(WE) foi proposta por Mudholkar et al.(1996) por meio de uma modificação da distribuição Weibull padronizada, em que foi acrescentado o parâmetro β. A função de densidade de probabilidade, que é caracterizada pelos parâmetros de forma $\alpha > 0$ e $\lambda > 0$, e pelo parâmetro de escala $\gamma > 0$, é expressa por:

$$f(t) = \frac{\alpha\lambda}{\gamma}\left[1 - \exp\left[-\left(\frac{t}{\gamma}\right)^{\alpha}\right]\right]^{\lambda-1} \exp\left[-\left(\frac{t}{\gamma}\right)^{\alpha}\right]\left(\frac{t}{\gamma}\right)^{\alpha-1}, t > 0.$$

As funções de sobrevivência e taxa de falha são, respectivamente,

$$S(t) = 1 - \left[1 - \exp\left[-\left(\frac{t}{\gamma}\right)^{\alpha}\right]\right]^{\gamma},$$

e

$$h(t) = \frac{\lambda\alpha(\frac{t}{\gamma})^{\gamma-1}\exp[-(\frac{t}{\gamma})^{\alpha}][1 - \exp(-(\frac{t}{\gamma})^{\alpha})]^{\lambda-1}}{1 - [1 - \exp(-(\frac{t}{\gamma})^{\alpha})]^{\lambda}}.$$

A função risco da WE pode assumir diversas formas como: constante, crescente, decrescente, unimodal e forma de banheira. A distribuição WE tem como casos particulares as distribuições Weibull (quando seu parâmetro de forma $\lambda = 1$) e exponencial (quando $\lambda = 1$ e $\gamma = 1$).

2.2.4 Distribuição Weibull modificada generalizada

A distribuição de probabilidade Weibull modificada generaliza (WMG) (Carrasco, 2007) pode ser usada para modelar diversos problemas em análise de sobrevivência devido à diversidade de formas que sua função risco pode assumir. Dada uma variável aleatória T contínua e não negativa, sua densidade de probabilidade é expressa pela função:

$$f(t) = \frac{\alpha\varphi(\gamma + \lambda t)t^{\gamma-1}\exp\{\lambda t - \alpha t^{\gamma}\exp(\lambda t)\}}{[1 - \exp\{-\alpha t^{\gamma}\exp(\lambda t)\}]^{1-\varphi}}.$$

As funções de distribuição acumulada, de sobrevivência e a função risco são dadas pelas fórmulas abaixo:

$$F(t) = [1 - \exp\{-\alpha t^{\gamma}\exp(\lambda t)\}]^{\varphi},$$

$$S(t) = 1 - [1 - \exp\{-\alpha t^{\gamma}\exp(\lambda t)\}]^{\varphi},$$

$$h(t) = \frac{\alpha\varphi(\gamma + \lambda t)t^{\gamma-1}\exp\{\lambda t - \alpha t^{\gamma}\exp(\lambda t)\}[1 - \exp\{-\alpha t^{\gamma}\exp(\lambda t)\}]^{\varphi-1}}{1 - [1 - \exp\{-\alpha t^{\gamma}\exp(\lambda t)\}]^{\varphi}}.$$

Sendo $\alpha \geq 0$ o parâmetro de escala, $\gamma \geq 0$ e $\varphi \geq 0$ parâmetros que modificam e determinam a forma da distribuição, chamados então de parâmetros de forma. Essa distribuição apresenta também um fator acelerador na sobrevida que é representado por $\exp(\lambda t)$, e $\lambda \geq 0$.

A função taxa de falha da distribuição WMG é flexível e pode assumir formas crescente, descrescente, constante, forma de U (também chamada de forma de banheira) e forma unimodal. Quando:

- $\gamma \geq 1$ e $\varphi < 1$, a função taxa de falha é crescente.
- $\gamma > 1$ e $0 < \varphi < 1$, a função risco tem forma de banheira.
- $< \gamma < 1$ e $0 < \varphi < 1$, a função taxa de falha tem forma de banheira.
- $\varphi \to \infty$, a função de risco assume uma forma unimodal.

Outra particularidade da distribuição de probabilidade WMG é que ela é uma generalização de diversas distribuições de probabilidade da família Weibull. Algumas das distribuições que são casos particulares da WMG são: exponencial, Weibull, exponencial exponencializada (Gupta e Kundu, 2001), Weibull exponenciada (Mudholkar et al. 1996) e Weibull modificada.

2.3 Estimação dos parâmetros

Existem diversos métodos para estimar o valor dos parâmetros dos modelos de probabilidade. O método de máxima verossimilhança se destaca na área de análise de sobrevivência, pois é um dos poucos que consegue incorporar em sua metodologia de cálculo a censura e possui propriedades ótimas para grandes amostras.

2.3.1 O método de Máxima verossimilhança

A estimação dos parâmetros no método de máxima verossimilhança (COLOSIMO e GIOLO, 2006) é feita baseada nos resultados obtidos da amostra, e a ideia desse método é achar a distribuição que tenha a maior probabilidade de ter gerado aquela amostra. Com esse intuito, o método procura os valores de $\boldsymbol{\theta}$ que maximizem a função de verossimilhança $L(\boldsymbol{\theta})$, pois são esses valores que têm a maior probabilidade de ter gerado a amostra analisada.

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta}).$$

Nesta expressão, n é o número total de observações e $\boldsymbol{\theta}$ representa o vetor de parâmetros que pode ter 1 ou mais parâmetros. Adequando o método de estimação para a análise de sobrevivência, define-se uma variável indicadora de censura δ_i que assumirá valor 0 quando a variável t_i representar um tempo de censura, e o valor 1 quando t_i for um tempo de falha. Cada observação da amostra contribui para a função de verossimilhança com $f(t_i, \boldsymbol{\theta})$ quando a variável indicadora de censura é 1, ou com $S(t_i, \boldsymbol{\theta})$ quando o tempo representa um tempo de censura.

Sendo assim, para calcular a função de sobrevivência as observações são divididas em dois conjuntos: as observações censuradas e as falhas. Deste modo, a função de verossimilhança levando em consideração a informação de censura é dada por:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} [f(t_i; \boldsymbol{\theta})_i^{\delta} [S(t_i; \boldsymbol{\theta}]^{1-\delta_i},$$

em que r é o número de censuras. Para obter os estimadores de máxima verossimilhança, primeiramente obtem-se com o logaritmo da função de verossimilhança:

$$\log(L(\boldsymbol{\theta})) = l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \log[f(t_i, \boldsymbol{\theta})] + (1 - \delta_i) \log[S(t_i, \boldsymbol{\theta})].$$

Logo após, deriva-se parcialmente a função $l(\boldsymbol{\theta})$ e resolve-se o sistema de equações abaixo:

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

2.3.2 Intervalo de Confiança e teste de hipótese dos Estimadores de Máxima Verossimilhança

Após estimar os parâmetros da distribuição pelo método de máxima verossimilhança, pode-se então calcular o intervalo de confiança dessas estimativas (COLOSIMO e GIOLO, 2006). Para a sua construção é necessário utilizar as propriedades assintóticas dos estimadores de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$. Uma das propriedades afirma que para grandes amostras, sob certas condições de regularidade, a distribuição do vetor $\widehat{\boldsymbol{\theta}}$ é normal multivariada com média $\boldsymbol{\theta}$ e matriz de variância-covariância $Var(\widehat{\theta})$. Dessa forma, a distribuição de $\widehat{\boldsymbol{\theta}}$ pode ser representada por:

$$\widehat{\boldsymbol{\theta}} \sim N_k(\boldsymbol{\theta}, Var(\widehat{\boldsymbol{\theta}})),$$

em que

$$Var(\widehat{\boldsymbol{\theta}}) \approx -[I_F(\boldsymbol{\theta})]^{-1},$$

e

$$I_F = E\left[\left(\frac{\partial \log L(\boldsymbol{\theta}}{\partial \boldsymbol{\theta}}\right)^2\right].$$

Devido à dificuldade de calcular a esperança, usa-se a matriz de informação observada avaliada em $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. Portanto, um intervalo de confiança aproximado de $(1-\alpha)100\%$ de confiança para $\boldsymbol{\theta}$ é dado por:

$$\widehat{\boldsymbol{\theta}} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{\boldsymbol{\theta}})}.$$

Após obter as estimativas e intervalos de confiança, é interessante testar hipóteses para o vetor de parâmetros(θ). Um das hipóteses que pode ser testada é :

$$H_0 : \boldsymbol{\theta}_j = \boldsymbol{\theta}_{j0}$$
$$H_1 : \boldsymbol{\theta}_j \neq \boldsymbol{\theta}_{j0}$$

A distribuição assintótica dos estimadores é a generalização da t de student, para grandes amostras a t se aproxima da distribuição normal.Portanto, o teste que é baseado na distribuição assintótica tem como estatística do teste:

$$Z = \frac{\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_{j0}}{\sqrt{\widehat{Var}(\widehat{\boldsymbol{\theta}}_j)}} \sim N(0,1).$$

2.4 Metologia de Fração de Cura

Os modelos de análise de sobrevivência assumem que todos os indivíduos na análise falharam ou irão falhar. Mas existem situações em que a função de sobrevivência não tende a zero quando o tempo tende a infinito, indicando que talvez exista uma fração de indivíduos que não irão falhar. Estes são considerados indivíduos curados. Os individuos curados são aqueles indivíduos que nunca experimentarão o evento de interesse, também chamados de imunes ou não suscetíveis (Berkson e Gage, 1952).

Historicamente, modelos que usam a metodologia de fração de cura são utilizados para estimar a fração de sobrevivência. Nos modelos de sobrevivência com fração de cura, a população estudada é dividida em dois grupos: em um deles estão os indivíduos suscestíveis e no outro os indivíduos curados. Suponha que exista uma variável indicadora associada a cada indivíduo i para indicar se o i-esimo indivíduo é curado ou suscetível ao evento de interesse, isto é,

$$V_i = \begin{cases} 1 & \text{se o indivíduo no } i\text{-ésimo evento é suscetível} \\ 0 & \text{se o indivíduo no } i\text{-ésimo evento é curado.} \end{cases}$$

Associada a essa variável indicadora temos uma probabilidade, em que $P(V_i = 1) = \phi$ representa a probabilidade do indivíduo ser imune e a probabilidade complementar $P(V_i = 0) = (1-\phi)$ é a probabilidade do indivíduo ser suscetível ao evento de interesse. Para incorporar a informação de fração de cura no modelo, reescreve-se a função de sobrevivência obtendo uma função de sobrevivência populacional:

$$S_{pop}(t) = (1-\phi) + \phi S(t). \tag{6}$$

As propriedades da função $S_{pop}(t)$ são: $\lim_{t\to\infty} S_{pop}(t) = (1-\phi)$ e $\lim_{t\to 0} S_{pop}(t) = 1$. A função densidade populacional é encontrada a partir da relação abaixo:

$$F_{pop}(t) = 1 - S_{pop}(t),$$

e de (1), tem-se que :

$$f_{pop}(t) = \frac{d}{dt}(1 - (1 - \phi) + \phi S(t)).$$

Portanto,

$$f_{pop}(t) = \phi f(t).$$

2.5 Modelos de Regressão

Os estudos em geral, na área de sobrevivência ou não, costumam buscar uma relação entre a variável resposta e as demais variáveis presentes em cada situação. Na área de saúde, por exemplo, o tempo até a morte de um paciente com câncer de pulmão, pode ser influenciado por algumas covariáveis como: idade, ser fumante ou não, tamanho do tumor e raça.

Uma das formas de relacionar a variável resposta, tempo, e as demais covariáveis é através de um modelo de regressão. Na análise de sobrevivência, existem duas classes importantes de modelos de regressão: os modelos de riscos proporcionais e os modelos de locação e escala (Carrasco, 2007).

Neste trabalho, será utilizado o modelo de locação e escala para estudar a influência das covariáveis.

2.5.1 Modelos de Locação e escala

A classe de modelos de regressão chamada modelos de locação e escala (Carrasco, 2007) tem uma característica importante: trabalha com o logaritmo do tempo de sobrevivência, isto é,

$$Y = \log(T).$$

O modelo de regressão de locação e escala tem a seguinte forma:

$$Y = \log(T) = \mu(\mathbf{x}) + \sigma \mathbf{z},$$

em que $\sigma > 0$ representa o parâmetro de escala e $\boldsymbol{z}$ é o erro aleatório. Geralmente, considera-se $\mu(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}$ em que $\boldsymbol{x}^T$ representa o vetor de covariáveis $\boldsymbol{x}^T = (x_1, x_2, ..., x_p)^T$ sendo p

a quantidade de covariáveis presente no modelo de regressão, portanto $\mu(\boldsymbol{x})$ é o parâmetro de locação dependente das variáveis regressoras e $\boldsymbol{\beta} = (\beta_1\beta_2, ..., \beta_p)$ representa o vetor de parâmetros desconhecidos. O modelo de locação e escala é portanto um modelo log linear para a variável T e as variáveis regressoras atuam multiplicativamente sobre T.

2.5.2 Modelos de Regressão Weibull

Seja T uma variável aleatória com distribuição Weibull e a variável aleatória $Y = \log(T)$ que tem distribuição do valor extremo com as reparametrizações $\alpha = \exp(\mu)$ e $\gamma = \frac{1}{\sigma}$ e função de densidade:

$$f(y) = \frac{1}{\sigma}\exp\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right).(7)$$

Sendo que y e $\mu \,\epsilon\, \Re$ e $\sigma > 0$. Os parâmetros σ e μ são respectivamente os parâmetros de escala e locação. Para incluir covariáveis no modelo $\mu = \mu(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ e o vetor $\boldsymbol{\beta}\epsilon\Re$. A função de sobrevivência é escrita como:

$$S(y) = \exp\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right).$$

2.5.3 Modelo de Regressão Weibull modificada

Seja T uma variável aleatória com distribuição Weibull modificada, o modelo de Regressão Weibull Modificada é encontrado por meio de uma substituição de variáveis, sendo esta $Y = \log(T)$. Faz-se necessário as reparametrizações $\alpha_1 = \log(\alpha)$, $\gamma = \frac{1}{\sigma}$ e $\lambda = \exp(-\mu)$ para encontrar a função densidade de probabilidade e a função de sobrevivência.

$$f(y) = \left[\frac{1}{\sigma} + \exp(y-\mu)\right]\exp\left[\alpha_1 + \left(\frac{y-\mu}{\sigma}\right) + \exp(y-\mu) + \frac{\mu}{\sigma}\right]\times$$
$$\exp\left[-\exp\left[\alpha_1 + \left(\frac{y-\mu}{\sigma}\right) + \exp(y-\mu) + \frac{\mu}{\sigma}\right]\right],$$

$$S(y) = \exp\left[-\exp\left[\alpha_1 + \left(\frac{y-\mu}{\sigma}\right) + \exp(y-\mu) + \frac{\mu}{\sigma}\right]\right].$$

A variável aleatória y e os parâmetros μ e $\alpha_1 \,\epsilon\, \Re$ e $\sigma > 0$. Para adicionar covariáveis ao modelo deve-se considerar o vetor de covariáveis $\boldsymbol{x}_T = (x_1, x2, ..., xp)^T$ e um vetor de parâmetros desconhecidos $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta p) \,\epsilon\, \Re$, então $\mu = \mu(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$.

2.5.4 Modelo de Regressão Weibull exponencializada

Se T é uma variável aleatória com distribuição Weibull Exponencializada, para obter um modelo de regressão para essa distribuição, usa-se $Y = \log(T)$, portanto a função distribuição de probabilidade de Y é dada por:

$$f(y) = \frac{\lambda}{\sigma}\exp\left(\left(\frac{y-\mu}{\sigma}\right) - exp\left(\frac{y-\mu}{\sigma}\right)\right)\left[1 - exp\left(-exp\left(\frac{y-\mu}{\sigma}\right)\right)\right]^{\lambda-1}, \qquad (8)$$

$$S(y) = 1 - \left\{1 - \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\lambda}, \qquad (9)$$

com $-\infty < y < \infty$, $\lambda > 0$, $\sigma > 0$ e $-\infty < \mu < \infty$, sendo σ e μ parâmetros de escala e locação respectivamente e λ parâmetro de forma. A maneira de adicionar covariáveis a esse modelo é a mesma descrita nos modelos anteriores.

2.5.5 Modelo de Regressão Weibull modificada generalizada

Seja T uma variável aleatória com distribuição Weibull Modificada Generalizada, a variável aleatória Y gerada pela transformação $Y = \log T$ tem distribuição log Weibull Modificada Generalizada com função distribuição e função de sobrevivência respectivamente:

$$f(y) = \frac{\varphi(\frac{1}{\sigma} + \lambda\exp(y))\exp[(\frac{y-\mu}{\sigma}) + \lambda\exp(y) - \exp[(\frac{y-\mu}{\sigma}) + \lambda\exp(y)]]}{[1-\exp[-\exp[(\frac{y-\mu}{\sigma}) + \lambda\exp(y)]]]^{1-\varphi}}, \qquad (10)$$

$$S(y) = 1 - \left[1 - \exp[-\exp[(\frac{y-\mu}{\sigma}) + \lambda\exp[\left(\frac{y-\mu}{\sigma}\right)\sigma]\exp(\mu)]]\right]^{\varphi}. \qquad (11)$$

As funções (10) e (11) são encontradas com as reparametrizações $\varphi = \frac{1}{\sigma}$ e $\alpha = \exp(\frac{-\mu}{\sigma})$ em que $-\infty < \mu < \infty$ e $-\infty < y < \infty$, $\sigma > 0$, $\varphi > 0$ e $\lambda \geq 0$. Os parâmetros φ e λ são parâmetros de forma, e os parâmetros μ e σ são respectivamente os parâmetros de locação e escala. Assim como nos demais modelos quando é útil a inclusão de covariáveis, $\mu = \mu(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ em que $\boldsymbol{x}_T = (x_1, x2, ..., xp)^T$ é o vetor de covariáveis e $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta p)\ \epsilon\ \Re$ é um vetor de parâmetros desconhecidos. O cálculo necessário para encontrar os modelos de regressão expostos nesse trabalho estão apresentados no Anexo.

3 METODOLOGIA

3.1 Material

No presente estudo será feita uma análise com dados de melano maligno (Ibraim at al. 2001). Esses dados são provenientes de um estudo com pacientes que foram operados de um câncer de pele (Melano cutâneo). O estudo tem como evento de interesse a reincidência do câncer, para a avaliação do desempenho pós-cirúrgico com alta dose da droga interferon alfa-2b.

O tempo do estudo foi de 1991 até 1998, porém o período de entrada dos pacientes foi de 1991 a 1995. A amostra é de 417 pacientes e a variável T mede o tempo até a morte do paciente. Cada paciente tem as seguintes variáveis associadas:

- tempo observado (em anos);
- indicador de censura (0=censurado, 1=falha);
- tipo de tratamento (0=observação, 1=interferon);
- idade (em anos);
- sexo (0=masculino, 1=feminino);
- performace status-escala de capacidade funcional do paciente quanto às suas atividades diárias (0=completamente ativo, 1=outro).
- espessura do tumor (em mm).

3.2 Métodos

Com base nas seções 2.4 e 2.5, o modelo estudado neste trabalho foi o Modelo de regressão Weibull modificada generalizada com fração de cura. Este modelo tem função de sobrevivência e função densidade de probabilidade representadas, respectivamente por:

$$S_{pop}(y) = (1-\phi) + \phi S(y),$$

em que S(y) é a sobrevivência do Modelo de regressão Weibull modificada generalizada dada pela equação (11). Logo tem-se:

$$S_{pop}(y) = (1-\phi) + \phi\left[1 - \left[1 - \exp[-\exp[(\frac{y-\mu}{\sigma}) + \lambda\exp[\left(\frac{y-\mu}{\sigma}\right)\sigma]\exp(\mu)\right]^{\gamma}\right] \qquad (12)$$

e

$$f_{pop} = \phi f(y),$$

em que f(y) é a função densidade de probabilidade do Modelo de regressão Weibull modificada generalizada dada pela equação (10), então:

$$f_{pop}(y) = \phi \left[\frac{\gamma[\sigma^{-1} + \lambda \exp{(y)}] \exp{(\frac{y-\mu}{\sigma})} + \lambda \exp{(y)} - exp[(\frac{y-\mu}{\sigma}] + \lambda \exp{(y)}}{[1 - exp[-exp((\frac{y-\mu}{)}\sigma) + \lambda \exp(y)]]^{1-\gamma}} \right], \tag{13}$$

em que $0 \leq \phi \leq 1$, $-\infty < \mu < \infty$ que representa o parâmetro de locação, pode ser escrito em função das covariáveis $\mu = x^T \beta$,$\sigma > 0$ é o parâmetro de escala , $\gamma > 0$ e $\lambda \geq 0$ são os parâmetros de forma. O modelo estudado tem como casos particulares:

- Modelo de regressão Weibull com fração de cura quando $\gamma = 1$ e $\lambda = 0$,
- Modelo de regressão Weibull Modificada com fração de cura quando $\gamma = 1$,
- Modelo de regressão Weibull Exponencializada com fração de cura quando $\gamma = 0$.

3.2.1 O método de Máxima verossimilhança com restrição nos parâmetros

O método de máxima verossimilhança com restrição no espaço paramétrico estima os valores que maximizam a função de verossimilhança. O vetor de parâmetros $\boldsymbol{\theta}$ tem k restrições de inequações lineares $\boldsymbol{u}_i\boldsymbol{\theta} - c_i \geq 0$, em que $\boldsymbol{u}_i$, i= 1,2,...,k, são k x 1 vetores e c_j são escalares que assumem valores 0 ou 1. A função de verossimilhança com restrição:

$$l(\boldsymbol{\theta}, v) = l(\boldsymbol{\theta}) + v\Sigma_{i=1}^{k}(\boldsymbol{u}_{i\boldsymbol{\theta}} - c_i)$$

o parâmetro de ajuste v é positivo. Neste estudo para obter $\widehat{\boldsymbol{\theta}}$ utilizou-se a função *constrOptim* do *software* estatistico R(R DEVELOPMENT CORE TEAM, 2012). Maiores detalhes sobre o método de máxima verossimilhança com restrição ver (Lange, 1999) e help do *software* R.

4 RESULTADOS

4.1 Análise descritiva dos dados

Com o intuito de conhecer os dados para uma análise posterior mais elaborada realizou-se uma análise exploratória dos dados. Primeiramente, constroi-se o gráfico da curva de Kaplan-Meier sem levar em consideração nenhuma covariável, representado na Figura 2.

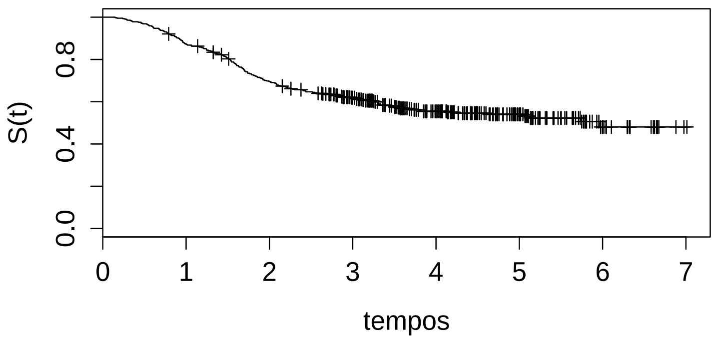

Figura 2 – Curva estimada pelo método não paramétrico de Kaplan-Meier para os tempos de sobrevivência de indivíduos com melano cutâneo

A Figura 2 mostra indícios de que a função de sobrevivência é imprópria, pois, quando $t \to \infty$, a função de sobrevivência não tende a zero, o que indica a existência de uma possível fração de cura entre os dados. Após a análise descritiva da função de sobrevivência estimada por Kaplan-Meier determinou-se a função taxa de falha empírica por meio da metodologia do gráfico do tempo total em teste (curva TTT), que ajuda na escolha de quais possíveis distribuições poderão ser considerados para analisar os dados de melanoma cutâneo.

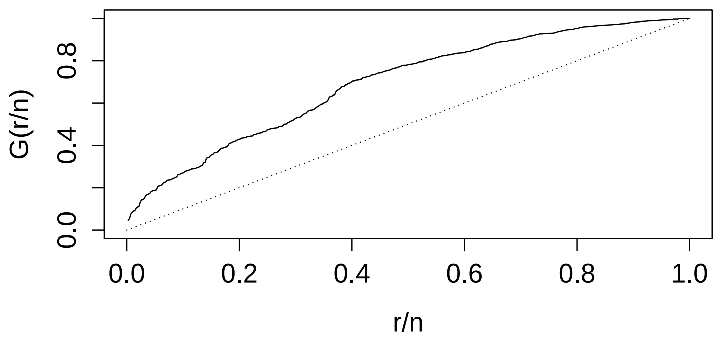

Figura 3 – Curva do tempo total em teste para os dados de paciêntes que foram operados do câncer melanoma cutâneo

A Figura (3) mostra que a curva TTT assume uma forma côncava, o que indica que a distribuição que modelará bem os dados será uma função distribuição de probabilidade que a função risco assuma forma estritamente crescente. Por esse motivo, escolheu-se as distribuições Weibull, Weibull generalizada, Weibull modificada generalizada e a Weibull exponencializada, para serem estudadas pois todas essas distribuições permitem a forma crescente da função taxa de falha. Após a análise descritivas de sobrevivência faz-se uma análise que leva em consideração a covariáveis presentes no banco de dados. As análises de Kaplan-Meier quando levado em consideração as covariáveis dão um indicativo da significância das variáveis no modelo de regressão. Por isso foi feita as curvas, os testes estatisticos e algumas medidadas descritivas para nortear quais variáveis possivelmente serão significativas no modelo.

Tipo do Tratamento	N	Percentual	Tempo máximo	Tempo mínimo	Tempo médio
Observação	204	48,9	6,98	0,15	3,14
Interferón	213	51,1	7,01	0,24	3,22
Total	417	100,0	-	-	-

Ao realizar o teste não paramétrico de Wilcoxon para verificar se as curvas de sobrevivência para cada tipo de tratamento(Observação e Interferón)são iguais, obteve-se (p-valor=0,76), portanto não há indícios para rejeita-se a hipótese nula de que as curvas são iguais. Como mostra a Tabela 1 os tempos máximo e mínimos não apresentam variações grandes quando é levada em consideração o tipo de tratamento.

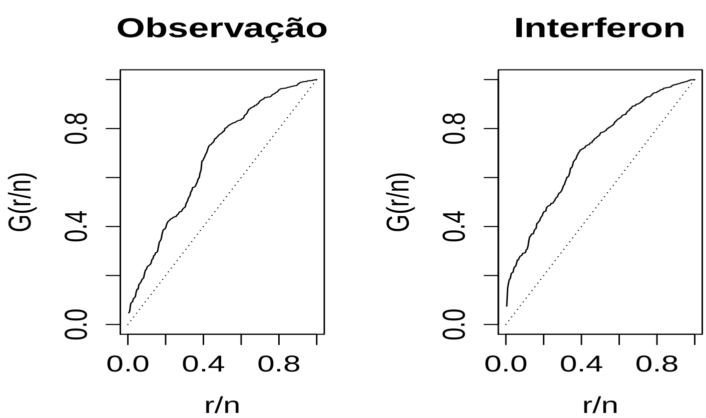

Figura 4 – Curvas estimadas pelo método não paramétrico de Kaplan-Meier para os tempos de sobrevivência de indivíduos com Melano Cutâneo por tipo do tratamento recebido

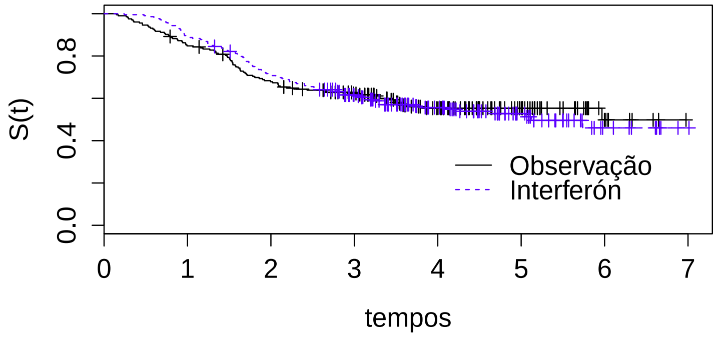

Figura 5 – Curvas estimadas pelo método não paramétrico de Kaplan-Meier para os tempos de sobrevivência de indivíduos com Melano Cutâneo por tipo do tratamento recebido

A Figura 4 mostra as curvas de sobrevivências não paramétricas estimadas pelo método de Kaplan-Meier pela covariável tipo de tratamento. A curva de sobrevivência não paramétrica dos indivíduos que receberam a droga (Interferón) não se diferencia muito da curva dos indivíduos que apenas foram observados, o que confirmam o teste não paramétrico de Wilcoxon de que as curvas não se diferenciam. As curvas TTT para os dois tipos de tratamentos apresentam o mesmo comportamento da Figura 3, que considera todos os tempos. Portanto, não existem indicativos de que os tempos se diferenciam devido a covariável tipo de tratamento.

Tabela 2 – Tabela descritiva da covariável sexo

Sexo	N	Percentual	Tempo máximo	Tempo mínimo	Tempo médio
Masculino	263	63,1	7,01	0,17	3,13
Feminino	154	36,9	6,65	0,15	3,27
Total	417	100,0	-	-	-

O estudo com a covariável sexo mostra que o tempo máximo e mínimo não apresentam variações grandes quando essa variável é levada em consideração.

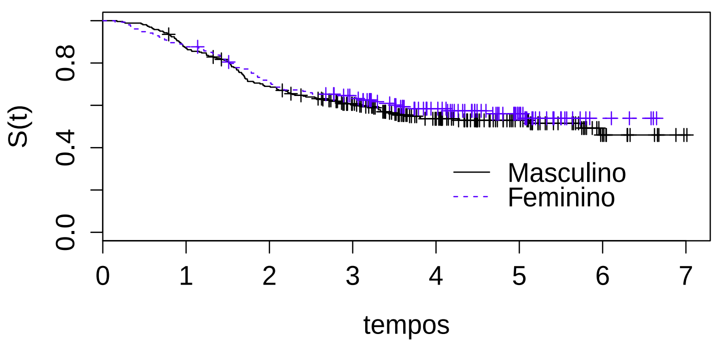

Figura 6 – Curva estimada pelo método não paramétrico de Kaplan-Meier para os tempos de sobrevivência de indivíduos com Melano Cutâneo por sexo

A Figura 6 expressa a curva de sobrevivência não paramétrica estratificada por sexo, pode-se observar que a curva de sobrevivência para o sexo feminino não se diferencia da curva de sobrevivência para o sexo masculino. Indicando que não existe diferença entre as curvas de sobrevivência quando é considerado a covariável sexo. O teste estatístico de Wilcoxon confirma a suspeita de que as curvas de sobrevivência não são diferentes, com o (p-valor=0,57) não existe evidências para rejeitar a hipótese nula de que as curvas são iguais, portanto não pode-se afirmar que as curvas são diferentes.

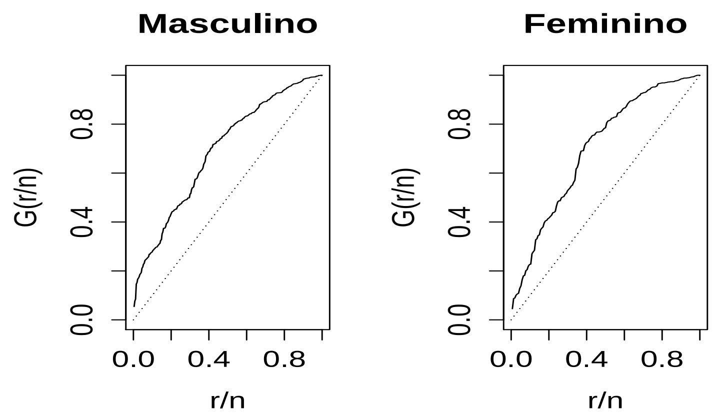

Figura 7 – Curva do tempo total em teste para os dados de sobrevivência de paciêntes que foram operados do câncer melanoma cutâneo pela variável sexo

A Figura 7 que expressa a curva do tempo total em teste mostra que o comportamento das curvas é o mesmo independente do sexo. Dessa forma, não existem evidências de que essa covariável será significativa no modelo.

Tabela 3 – Tabela descritiva da covariável Performace Status-Escala

Performace Status-Escala	N	Percentual	Tempo máximo	Tempo mínimo	Tempo médio
Completamente Ativo	363	87,1	7,01	0,15	3,22
Outro	54	12,9	6,58	0,17	2,94
Total	417	100,0	-	-	-

A covariável Performace Status-Escala apresenta uma quantidade maior de indivíduos na categoria completamente ativo, mas o tempo não parece se diferenciar nas categorias. Os testes estatísticos de Log-rank (p-valor = 0,255) e de Wilcoxon(p-valor=0,198) não são significativos portanto, não há evidências para afirmar que as curvas são diferentes.

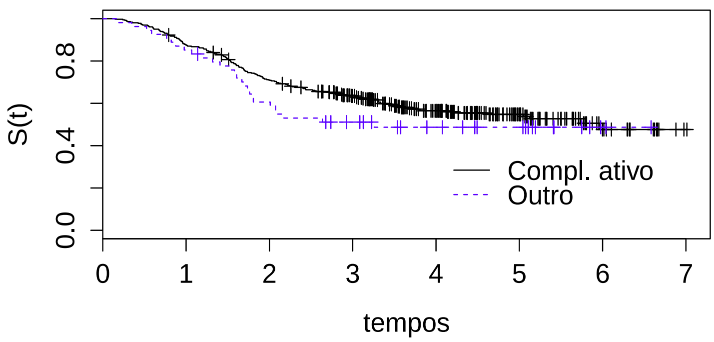

Figura 8 – Curva estimada pelo método não paramétrico de Kaplan-Meier para os tempos de sobrevivência de indivíduos por performace Status escala de capacidade funcional do paciente quanto ás suas atividades diárias

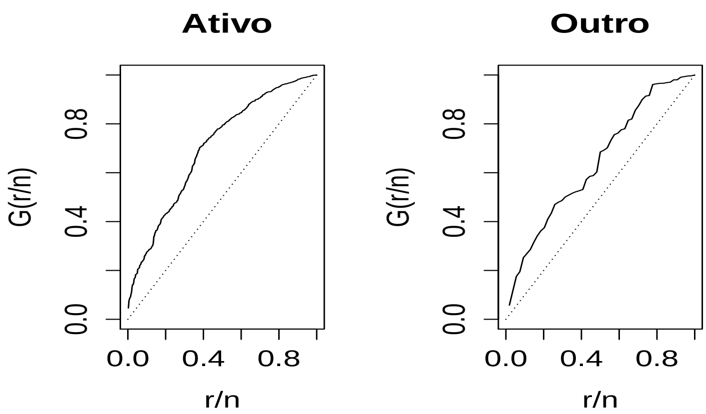

Figura 9 – Curva do tempo total em teste para os dados de sobrevivência de paciêntes que foram operados do câncer melanoma cutâneo pela performace status-escala

Na Figura 8 a curva de sobrevivência não paramétrica dos indivíduos completamente ativos para suas funções diárias parece se diferenciar no intervalo de tempo entre 2 a 3 anos e meio da curva de sobrevivência dos indivíduos com uma performace status escala igual a outro. Apesar dos testes não paramétricos não confirmarem essa diferença. Apesar das curvas total em teste expressas na Figura 9 serem estritamente crescentes, o crescimento da curva TTT para os indivíduos na categoria completamente ativo é maior do que a categoria da outro.

Tabela 4 – Tabela descritiva da variáveis quantitativas

Variável	Mínimo	Máximo	Média	Mediana
Tempo (em anos)	0,15	7,01	3,18	3,22
Idade (em anos)	19,13	78,05	57,56	48,00
Espessura do tumor (em mm)	0,08	20,00	5,10	3,94

A Tabela acima mostra as estatísticas descritivas das variáveis quantitativas presentes no banco de dados.

4.2 Modelagem

Após a análise descritiva, foi feita uma modelagem das distribuições de probabilidade (Weibull, Weibull Exponencializada, Weibull Modificada, Weibull Modificada Generalizada) e como mostra a Figura (10), a distribuição que parece melhor se adequar aos dados é a Weibull Exponencializada.

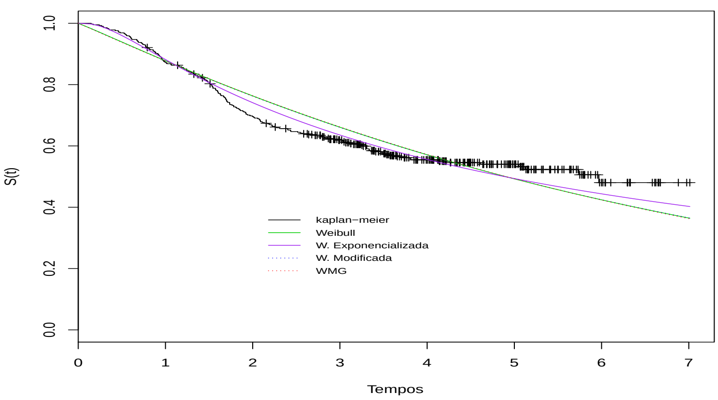

Figura 10 – Ajuste da curva de Kaplan-Meier e das curvas de sobrevivência estimada aos dados de melanôma cutâneo

Tabela 5 – Estimativas do critério de AIC para as distribuições da família Weibull

Modelo	AIC
WMG	1085,64
Weibull	1102,20
Weibull Modificada	1069,12
Weibull exponencializada	1068,96

Confirmando a análise gráfica o critério de AKAIKE (AIC), mostra que o melhor modelo é com a distribuição Weibull Exponencializada, pois esse modelo tem o menor AIC. Portanto, estimou-se o modelo Weibull Exponencializada com fração de cura e com as covariáveis presentes no banco de dados. As estimativas dos parâmetros, erro padrão e p-valor do modelo de regressão Weibull exponencializada com fração de cura encontra-se na Tabela 6.

Tabela 6 – Estimativas de Máxima Verossimilhança para o modelo Weibull exponencializada com fração de cura

Parâmetros	Estimativas	Erro Padrão	P-valor
σ	2,109	1,21	-
λ	13,867	21,66	-
β_0	-1,795	2,36	0,4474
β_1	0,279	0,14	0.0494
β_2	-0,004	0,005	0.4569
β_3	-0,113	0,14	0.4262
β_4	-0,188	0,19	0.3157
β_5	0,035	0,019	0.0695
ϕ	0,525	0,045	-

A Tabela 6 mostra que a proporção de cura estimada foi de $\widehat{1-p} = 0,475$ e os parâmetros significativos a um nível de significância de 5% foi o β_1 que acompanha a covariável tipo do tratamento. Porém, a um nível de significância de 10% o parâmetro β_5 que acompanha a covariável Diâmetro do tumor também é significativa. A interpretação do modelo é que os indivíduos que foram submetidos ao tratamento com a droga Interferon tem uma probabilidade maior de sobreviver ao tempo t, assim como a medida que a espessura do tumor aumenta, diminui a probabilidade de sobreviver ao tempo t. Foram feitos testes da razão de máxima verossimilhança comparando os modelos quando tira-se as covariáveis uma por uma e no modelo final as covariáveis que continuaram significativa foram a espessura do tumor (β_5)e o tipo de tratamento (β_1).

Tabela 7 – Estimativas de Máxima Verossimilhança para o odelo Weibull exponencializada com fração de cura

Parâmetros	Estimativas	Erro Padrão	P-valor
σ	1,981	0,961	-
λ	11,573	14,611	-
β_0	-1,790	1,865	0,337
β_1	0,298	0,133	0,026
β_5	0,032	0,0190	0,093
ϕ	0,519	0,042	-

As estimativas dos parâmetros do modelo Weibull exponencializada com fração de cura considerando as covariáveis Tipo de tratamento e espessura do tumor expressas na Tabela 7 mostram que os parâmetros foram significativos a um nível de 10%, menos o intercepto. Conclui-se portanto que os indivíduos submetidos ao tratamento tem uma probabilidade maior de sobreviver ao tempo t e a medida que a espessura do tumor aumenta, o individuo aumenta a probabilidade de sobreviver ao tempo t. A função densidade do modelo final que considera as duas covariáveis significativas é definida como:

$$f_{pop}(y) = \phi\Bigg[\frac{y}{\sigma}\exp\left[\left(\frac{y-(\beta_0+\beta_1x_1+\beta_2x_5)}{\sigma}\right) - \exp\left(\frac{y-(\beta_0+\beta_1x_1+\beta_2x_5)}{\sigma}\right)\right] \\ X\left[1-\exp\left(-\exp\left(\frac{y-(\beta_0+\beta_1x_1+\beta_2x_5)}{\sigma}\right)\right)\right]^{\gamma-1}\Bigg]$$

e a função de sobrevivência é escrita como,

$$S_{pop}(y) = (1-\phi) + (\phi)\left[1-\left[1-\exp\left[-\exp\left(\frac{y-(\beta_0+\beta_1x_1+\beta_2x_5)}{\sigma}\right)\right]\right]^{\gamma}\right]$$

Com o intuito de investigar a significância do parâmetro de indivíduos curados,(1 - ϕ), utilizou-se o teste de hipótese. Segundo Maller e Zhow(1996) para testar H_0: ϕ

$= 1$(não existe presença de indivíduos curados) deve-se considerar o fato que o teste de hipótese está considerando a fronteira do espaço paramétrico associado a esse parâmetro. Dessa forma, a estatística deviance definida por, $dn = 2(ln(\widehat{\theta_n}) - ln(\widehat{\theta_{H_0}}))$, tem aproximadamente distribuição χ_1^2 definida por $P(X \leq x) = \frac{1}{2} + \frac{1}{2}P(\chi_1^2 \leq x)$. O percentil de 95%, $C_{0.95}$, dessa distribuição é de $C_{0.95}$= 2,71. O valor da estatística deviance para os modelos Weibull exponencializada com fração de cura e Weibull exponencializada sem fração de cura é de dn= 32,93. Sendo assim, como $32,93 > 2,71$, há evidências de existência de indivíduos curados na amostra observada, e a estimativa do parâmetro de cura(1-$\widehat{\phi}$)= 0,481 é alta.

REFERÊNCIAS

AARSET, M.V. How to identify bathtub hazard rate. **IEEE Transactions on Reliability**, New York, v.36, p.106-108,1987.

BERKSON, J.; GAGE, R.P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, v.47, p. 501-515, 1952.

CARRASCO, J.M.F. **Modelos de Regressão log-Weibull modificado e a nova distribuição Weibull modificada generalizada**. 2007. 129p. Dissertação (Mestrado em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2007.

COLOSIMO, E.A.; GIOLO, S.R. **Análise de sobrevivência aplicada**, São Paulo: Blucher, 2006. 392p.

FACHINI, J.B **Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência**. 2011. 140p. Dissertação (Doutorado em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2011.

GUPTA, R.D.; KUNDU, D. Exponentiated exponential family: An alternative to Gamma and Weibull distributions. **Biometrical Journal**, Berlin, v.43, p. 117-130, 2001.

IBRAHIM, J. G.; CHEN, M. H.; SINHA, D. **Bayesian survival analysis**, New York: Springer, p.479, 2001.

KALBFLEISCH, J.D.; PRENTICE, R.L. **The statistical analysis of failure time data**. 2nd ed. New York: John Wiley, 2002. 439p.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival analysis: techniques for censored and truncated data**. New York: Springer Verlag, 1997. 536 p.

LAI C.D., XIE, M.; MURTHY, D.N.P. A modified Weibull Distribution. **IEEE Transactions on Reliability**, New York, v.52, p.33-37, 2003.

LANGE, K. **Numerical analysis for statisticians.** New York: Springer, 1999. 356 p.

LAWLESS, J. F. **Statistical models and methods for lifetime data**, 2nd ed., New York: Wiley, 2003. 439 p.

MAGALHÃES, M.N. **Probabilidade e Variáveis Aleatórias**, São Paulo: Editora da Universidade de São Paulo, 2011.

MALLER,R.; ZHOU, X. **Survival Analysis with long-Term Survivors**, John Wiley Sons, 1996.

MUDHOLKAR, G.S.; SRIVASTAVA, D.K. Exponentiated Weibull Family for Analyzing Bathtub Failure-Rate Data. **IEEE Transactions on Reliability**, New York, v.42, p.229 - 302, 1993.

MUDHOLKAR, G.S.; SRIVASTAVA, D.K.; FREIMER, M. The exponentiated Weibull family a reanalysis of the bus-motor-failure data. **Technometrics**, v.37,p. 436-445, 1995.

MUDHOLKAR, G.S.; SRIVASTAVA, D.K.; KOLLIA, G.D. A generalization of the Weibull Distribution with Application to the Analysis of Survival Data. **Journal of the American Statistical Association**, Boston, v.91,p. 1575- 1583, 1996.

MURTHY,D.N.P.; XIE, M.; JIANG, R. **Weibull models**, Wiley series in probability and statistics, New Jersey, Canada, 2004.

R Core Team (2012). **R**:A language and environment for statistical computing, Vienna, Austria, Disponível em: http://www.R-project.org/.

ANEXOS

Anexo A: Modelos de regressão locação e escala

Modelo Valor Extremo

Considere t uma variável aleatória com distribuição Weibull. Seja $Y = log(t)$. Tem-se, então, que:

1. $t = exp(y)$

2. $f_Y(y) = f_T(y)|J|$, em que:

   $|J| = \frac{dt}{dy} = exp(y)$

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left(- \left(\frac{t}{\alpha}\right)^\gamma \right)$$

Portanto,

$$\begin{aligned} f_Y(y) &= \frac{\gamma}{\alpha^\gamma} \exp{(y)}^{\gamma-1} \exp\left[- \left(\frac{\exp(y)}{\alpha}\right)^\gamma \right] \exp(y) \\ &= \frac{\gamma}{\alpha^\gamma} \exp{(y)}^{\gamma} \exp\left[- \left(\frac{\exp(y)}{\alpha}\right)^\gamma \right] \\ &= \frac{c}{s^c} exp(cy) \left[1 + \left(\frac{exp(cy)}{s^c}\right) \right]^{-2}. \end{aligned}$$

Fazendo as seguintes substituições: $\gamma = \frac{1}{\sigma}$ e $\alpha = \exp(\mu)$, tem-se que:

$$\begin{aligned} f_Y(y) &= \frac{1}{\sigma \exp(\mu)^{\frac{1}{\sigma}}} \exp{(y)}^{\frac{1}{\sigma}} \exp\left[- \left(\frac{\exp(y)}{\exp(\mu)}\right)^{\frac{1}{\sigma}} \right] \\ &= \frac{1}{\sigma} \exp\left(\frac{y-\mu}{\sigma}\right) \exp\left[- \left(\exp\left(\frac{y-\mu}{\sigma}\right)\right) \right] \\ &= \frac{1}{\sigma} \exp\left[\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right) \right]. \end{aligned}$$

Portanto, $Y \sim ValorExtremo(\mu, \sigma)$.

Para $S(y)$ tem-se $S(y) = 1 - F(y)$.

Modelo Log-Weibull Modificada

Considere t uma variável aleatória com distribuição Weibull modificada. Seja $Y = log(t)$. Tem-se, então, que:

1. $t = exp(y)$

2. $f_Y(y) = f_T(y)|J|$, em que:

   $|J| = \frac{dt}{dy} = exp(y)$

$$f(t) = \alpha(\gamma + \lambda t)t^{\gamma-1}\exp(\lambda t)\exp\left[- \alpha t^{\gamma}\exp(\lambda t)\right]$$

Portanto,

$$f_Y(y) = \alpha(\gamma + \lambda\exp(y))\exp(y\gamma)\exp(-y)\exp(\lambda\exp(y))\exp[-\alpha\exp(y\gamma)\exp(\lambda\exp(y))]\exp(y).$$

Fazendo as seguintes substituições: $\alpha_1 = \log(\alpha)$, $\lambda = \exp(-\mu)$ e $\gamma = \frac{1}{\sigma}$, tem-se que:

$$\begin{aligned}
f_Y(y) &= \exp(\alpha)\left(\frac{1}{\sigma} + \exp(y)\exp(-\mu)\right)\exp\left(\frac{y}{\sigma}\right)\exp(\exp(y-\mu))\exp[-\exp(\alpha)\exp\left(\frac{y}{\sigma})\exp(y-\mu)\right)] \\
&= \left(\frac{1}{\sigma} + \exp(y-\mu)\right)\exp\left(\frac{y}{\sigma} + \alpha_1)\exp(\exp(y-\mu)\right)\exp\left[-\exp\left(\frac{y}{\sigma} + \alpha_1 + (y-\mu)\right)\right] \\
&= \left(\frac{1}{\sigma} + \exp(y-\mu)\right)\exp\left(\frac{y}{\sigma} + \alpha_1\right)\exp(\exp(y-\mu))\exp\left[-\exp\left(\frac{y}{\sigma} + \alpha_1 + (y-\mu)\right)\right] \\
&= \left(\frac{1}{\sigma} + \exp(y-\mu)\right)\exp\left(\frac{y}{\sigma} + \alpha_1 + \exp(y-\mu)\right)\exp\left[-\exp\left[\frac{y}{\sigma} + \alpha_1 + (y-\mu)\right]\right] \\
&= (\frac{1}{\sigma} + \exp(y-\mu))\exp(\frac{y-\mu}{\sigma} + \alpha_1 + \exp(y-\mu) + \frac{\mu}{\sigma})\exp\left[-\exp\left[\frac{y-\mu}{\sigma} + \alpha_1 + (y-\mu) + \frac{\mu}{\sigma}\right]\right].
\end{aligned}$$

Portanto, $Y \sim log - WeibullModificada(\mu, \sigma, \alpha_1)$.

Para $S(y)$ tem-se que $S(y) = 1 - F(y)$.

Modelo Log-Weibull Exponencializada

Considere t uma variável aleatória com distribuição Weibull Exponencializada. Seja $Y = log(t)$. Tem-se, então, que:

1. $t = exp(y)$

2. $f_Y(y) = f_T(y)|J|$, em que:

   $|J| = \frac{dt}{dy} = exp(y)$

$$f(t)=\frac{\alpha\lambda}{\gamma}\left[1-\exp\left[-\left(\frac{t}{\gamma}\right)^{\alpha}\right]\right]^{\lambda-1}\exp\left[-\left(\frac{t}{\gamma}\right)^{\alpha}\right]\left(\frac{t}{\gamma}\right)^{\alpha-1}$$

Portanto,

$$f_Y(y)=\frac{\alpha\lambda}{\gamma}\left[1-\exp\left[-\left(\frac{\exp(y)}{\gamma}\right)^{\alpha}\right]\right]^{\lambda-1}\exp\left[-\left(\frac{\exp(y)}{\gamma}\right)^{\alpha}\right]\left(\frac{\exp(y)}{\gamma}\right)^{\alpha-1}(\exp(y))$$

Fazendo as seguintes substituições: $\alpha=\frac{1}{\sigma}$ e $\mu=\log(\gamma)$, tem-se que:

$$\begin{aligned}
f_Y(y) &= \frac{\lambda}{\sigma}\big[1-\exp\big[-\big(\frac{\exp(y)}{\exp(\mu)}\big)^{\frac{1}{\sigma}}\big]\big]^{\lambda-1}\exp\big[\big(\frac{\exp(y)}{\exp(\mu)}\big)^{\frac{1}{\sigma}}\big]\big(\frac{\exp(y)}{\exp(\mu)}\big)^{\frac{1}{\sigma}}\\
&= \frac{\lambda}{\sigma}\big(\exp(\frac{y-\mu}{\sigma})\big)\exp\big[-\big[\exp\big(\frac{y-\mu}{\sigma}\big)\big]\big]\big[1-\exp\big[-\exp(\frac{y-\mu}{\sigma})\big]\big]^{\lambda-1}\\
&= \frac{\lambda}{\sigma}\exp\big[\big(\frac{y-\mu}{\sigma}\big)-\exp\big(\frac{y-\mu}{\sigma}\big)\big]\big[1-\exp\big[-\exp\big(\frac{y-\mu}{\sigma}\big)\big]\big]^{\lambda-1}.
\end{aligned}$$

Portanto, $Y\sim log-Weibullexponencializada(\mu,\sigma,\lambda)$.

Para $S(y)$ tem-se $S(y)=1-F(y)$.

Modelo Log-Weibull modificada generalizada

Considere t uma variável aleatória com distribuição Weibull modificada generalizada. Seja $Y=log(t)$. Tem-se, então, que:

1. $t=exp(y)$

2. $f_Y(y)=f_T(y)|J|$, em que:

   $|J|=\frac{dt}{dy}=exp(y)$

$$f(t)=\frac{\alpha\varphi(\gamma+\lambda t)t^{\gamma-1}\exp\{\lambda t-\alpha t^{\gamma}\exp(\lambda t)\}}{[1-\exp\{-\alpha t^{\gamma}\exp(\lambda t)\}]^{1-\varphi}}.$$

Portanto,

$$f_Y(y)=\frac{\alpha\varphi(\gamma+\lambda\exp(y))\exp(y)^{\gamma}\exp[\lambda\exp(y)-\alpha\exp(y)^{\gamma}\exp(\lambda\exp(y))]}{1-\exp[-\alpha\exp(y)^{\gamma}\exp(\lambda\exp(y))]^{1-\varphi}}$$

Fazendo as seguintes substituições: $\gamma=\frac{1}{\sigma}$ e $\alpha=\exp(\frac{-\mu}{\sigma})$, tem-se que:

$$\begin{aligned} f_Y(y) &= \frac{\exp(\frac{\mu}{\sigma})\varphi(\frac{1}{\sigma}+\lambda\exp(y))\exp(\frac{y}{\sigma})\exp[\lambda\exp(y)-\exp(\frac{y-\mu}{\sigma})\exp(\lambda\exp(y))]}{[1-\exp[-\exp(\frac{y-\mu}{\sigma})\exp(\lambda\exp(y))]]^{1-\varphi}} \\ &= \frac{\varphi(\frac{1}{\sigma}+\lambda\exp(y))\exp(\frac{y-\mu}{\sigma})\exp[\lambda\exp(y)-\exp((\frac{y-\mu}{\sigma})+\lambda\exp(y))]}{[1-\exp[-\exp[(\frac{y-\mu}{\sigma})\exp(\lambda\exp(y))]]]^{1-\varphi}} \\ &= \frac{\varphi(\frac{1}{\sigma}+\lambda\exp(y))\exp[(\frac{y-\mu}{\sigma})+\lambda\exp(y)-\exp[(\frac{y-\mu}{\sigma})+\lambda\exp(y)]]}{[1-\exp[-\exp[(\frac{y-\mu}{\sigma})+\lambda\exp(y)]]]^{1-\varphi}}. \end{aligned}$$

Portanto, $Y \sim log - Weibullmodificadageneralizada(\mu, \sigma, \lambda, \varphi)$.

Para $S(y)$ tem-se $S(y) = 1 - F(y)$.